Journal of Agricultural Science (2002), **139**, 67–77. © 2002 *Cambridge University Press*
DOI: 10.1017/S0021859602002125 *Printed in the United Kingdom*

Two-stage clustering in genotype-by-environment analyses with missing data

A. J. R. GODFREY[1]*, G. R. WOOD[1], S. GANESALINGAM[1], M. A. NICHOLS[2] AND C. G. QIAO[1]

[1] *Institute of Information Sciences and Technology, Massey University, Palmerston North, New Zealand*
[2] *Institute of Natural Resources, Massey University, Palmerston North, New Zealand*

(*Revised MS received 14 March 2002*)

SUMMARY

Cluster analysis has been commonly used in genotype-by-environment (G × E) analyses, but current methods are inadequate when the data matrix is incomplete. This paper proposes a new method, referred to as two-stage clustering, which relies on a partitioning of squared Euclidean distance into two independent components, the G × E interaction and the genotype main effect. These components are used in the first and second stages of clustering respectively. Two-stage clustering forms the basis for imputing missing values in the G × E matrix, so that a more complete data array is available for other G × E analyses. Imputation for a given genotype uses information from genotypes with similar interaction profiles. This imputation method is shown to improve on an existing nearest cluster method that confounds the G × E interaction and the genotype main effect.

INTRODUCTION

Multi-environment trials are conducted for crop improvement and selection to examine the genotype-by-environment (G × E) interaction. The importance of interaction effects in G × E analyses has been well documented over the last 30 years. Review papers by Freeman (1973, 1985), Lin *et al.* (1986), Crossa (1990), as well as Cooper and DeLacy (1994) identify many of the models used in G × E analyses and explain their inter-relationships. Cluster analysis is used in G × E analyses to identify distinct groups of homogeneous genotypes or environments. In such studies, clustering is performed on the basis of both genotype main effect and G × E interaction (e.g. Mungomery *et al.* 1974), or the G × E interaction alone (e.g. Lin 1982).

Researchers are frequently faced with the problem of analysing incomplete and often unbalanced G × E matrices which arise as multi-environment trials progress over seasons, with genotypes being added or deleted. Removal of some data to form a complete G × E matrix wastes information and is consequently undesirable. For this reason, if clustering is to be applied to incomplete data sets, one of two routes needs to be taken: either the clustering procedure must be modified to handle the missing data, or the missing entries must be imputed so that standard cluster analysis can be performed. We pursue the first route and then use the modified clustering as the basis of imputation, thus providing an alternative method to that suggested by Drake (1981). Drake (1981) used the information from the closest cluster of genotypes to impute missing values in the G × E matrix. This method, however, confounds the genotype main effect and G × E interaction; these effects must be separated so that improved comparisons can be made. A two-stage method for clustering is developed in this paper. It uses a G × E interaction-based distance measure in the first stage, while in the second stage a main effect-based distance measure is used to find subclusters within first stage clusters.

The contributions of this paper are presented in five sections. The next section describes the partitioning of squared Euclidean distance and develops the relevant distance measures; this is followed by a section which describes two-stage clustering. A further section shows how realistic estimates of missing values can be imputed using results from two-stage clustering, and compares this new imputation method with existing methods. A well-known data set from the G × E literature (Mungomery *et al.* 1974) is then used to illustrate the proposed clustering and imputation methods. Our imputation results are then comprehensively compared with those found using an improvement we propose of the method due to Drake (1981).

* To whom all correspondence should be addressed. Email: A.J.Godfrey@massey.ac.nz

MAIN EFFECT AND INTERACTION DISTANCES

The familiar relationship

$$\sum_{k=1}^{K} y_k^2 = K\bar{y}^2 + (K-1)\frac{\sum_{k=1}^{K}(y_k-\bar{y})^2}{K-1}$$

partitions the sum of squares of observations $y_1, \ldots, y_k$ into orthogonal, and hence independent components, the first related to the level or the sample mean and the second to the variability or the sample variance.

Let y_{ik} be the yield of the ith genotype in the kth environment. Substituting the difference $y_{ik}-y_{jk}$ of two genotypes in the kth environment for y_k gives

$$\sum_{k=1}^{K}(y_{ik}-y_{jk})^2 = K(\bar{y}_{i\cdot}-\bar{y}_{j\cdot})^2 + (K-1)\frac{\sum_{k=1}^{K}((y_{ik}-y_{jk})-(\bar{y}_{i\cdot}-\bar{y}_{j\cdot}))^2}{K-1}$$

where $\bar{y}_{i\cdot}$ and $\bar{y}_{j\cdot}$ are the means of y_{ik} and y_{jk} respectively across all K environments. The left-hand-side of the last equation is the well-known expression for squared Euclidean distance

$$E_{ij}^2 = \sum_{k=1}^{K}(y_{ik}-y_{jk})^2,$$

as used in G × E analyses (Mungomery *et al.* 1974), while the right-hand-side exposes two components: a measure of the difference in genotype means and a measure of the G × E interaction.

The G × E interaction term has been used as a distance measure in clustering genotypes in the past. It has been commonly found by adjusting each row of the G × E matrix by the genotype mean, so that the distance is formed by summing squared differences in the centred rows of the G × E matrix. That is, using

$$D_{ij}^2 = \sum_{k=1}^{K}((y_{ik}-\bar{y}_{i\cdot})-(y_{jk}-\bar{y}_{j\cdot}))^2 \qquad (1)$$

Variants of this measure have been proposed. For example, the correlation between a pair of genotypes, and the scaling of all distances found using (1) by $2/(K-1)$ (Lin 1982) have been used. Lin (1982) stated that the adjusted rows of the G × E matrix are indicative of the 'shape' of a genotype's performance across environments, and termed the mean yield of each genotype its 'level'. Differences in these shapes or profiles will indicate the existence of G × E interaction.

Letting $I_{ij}^2 = D_{ij}^2/(K-1)$ denote the squared interaction distance, we can partition the squared Euclidean distance as

$$E_{ij}^2 = K(\bar{y}_{i\cdot}-\bar{y}_{j\cdot})^2 + (K-1)I_{ij}^2$$

This expression is now modified to allow for missing values in a G × E matrix.

Missing values in the G × E matrix will make the calculation of squared Euclidean distance possible only over environments in which both genotypes were grown. Ouyang *et al.* (1995) proposed dividing the sum of squared differences by the number p_{ij} of differences available, so that distances are not adversely affected by the number of common environments in which the pair of genotypes are grown, to give the expression

$$\frac{\sum_{\text{common } k}(y_{ik}-y_{jk})^2}{p_{ij}} \qquad (2)$$

for a mean squared Euclidean distance.

Care is needed when centring the rows of G × E matrices with missing values, as now genotype means will generally be based on distinct sets of environments, rendering the orthogonal decomposition invalid. To remedy this, for each pair of genotypes i and j calculate genotype means using only values from the p_{ij} environments for which both y_{ik} and y_{jk} have been recorded. The orthogonal partition of the squared Euclidean distance for rows i and j is then

$$E_{ij}^2 = p_{ij}(\bar{y}_{i\cdot}^{(j)}-\bar{y}_{j\cdot}^{(i)})^2 + (p_{ij}-1)\frac{\sum_{\text{common } k}((y_{ik}-\bar{y}_{i\cdot}^{(j)})-(y_{jk}-\bar{y}_{j\cdot}^{(i)}))^2}{p_{ij}-1}$$

with $\bar{y}_{i\cdot}^{(j)}$ defined as the mean of the yields y_{ik}, using only the p_{ij} environments common to both genotypes i and j. The means used for row-centring genotypes and the partitioning of squared Euclidean distance are therefore dependent on the particular pair of genotypes that are being compared. The partition of squared Euclidean distance, when data are incomplete, uses p_{ij} in place of the K (the total number of environments) used in the complete data case given earlier.

Godfrey *et al.* (1999) presented a distance measure based on the difference in levels of genotypes in a G × E matrix, referred to as the 'main effect distance', M_{ij}, where

$$M_{ij} = \frac{\left|\sum_{\text{common } k}(y_{ik}-y_{jk})\right|}{p_{ij}} \qquad (3)$$

which is equal to $|\bar{y}_{i\cdot}^{(j)}-\bar{y}_{j\cdot}^{(i)}|$. Use of the main effect distance recognizes that the comparison of means as a measure of the difference in level of a pair of genotypes is not valid when some data are missing.

The partition of squared Euclidean distance can now be expressed completely in terms of main effect and interaction distances as

$$E_{ij}^2 = p_{ij}M_{ij}^2 + (p_{ij}-1)I_{ij}^2$$

where the interaction distance I_{ij} is now given by

$$I_{ij} = \sqrt{\frac{\sum_{\text{common } k} ((y_{ik}-\bar{y}_{i\cdot}^{(j)})-(y_{jk}-\bar{y}_{j\cdot}^{(i)}))^2}{p_{ij}-1}} \qquad (4)$$

This distance expression, which measures G × E interaction differences among genotypes, is appropriate when there are missing entries in the G × E matrix. Its construction takes two ideas into account: the value used for row-centring a given genotype is tailored to the other genotype in the pair, and as we now show, appropriate averaging is used.

We now show how the distance measures I_{ij} and M_{ij} relate to a two-way model for the data. Such a model, assuming no replication in a cell, is

$$Y_{ik} = \mu + G_i + E_k + (GE_{ik} + \epsilon_{ik})$$

where the ϵ_{ik} are independent and normally distributed, with mean zero and variance σ^2. Note that the G × E interaction GE_{ik} and error ϵ_{ik} are confounded when there is no replication. Assuming that there are no missing data, the squared Euclidean distance between genotypes i and j is

$$E_{ij}^2 = \sum_{k=1}^{K} (\hat{G}_i - \hat{G}_j + \widehat{GE}_{ik} - \widehat{GE}_{jk} + e_{ik} - e_{jk})^2,$$

and therefore combines the genotype main effect and the difference in G × E interaction. Here a 'hat' denotes an estimator; note that $\bar{y}_{i\cdot} = \hat{\mu} + \hat{G}_i$. Also,

$$D_{ij}^2 = \sum_{k=1}^{K} ((\widehat{GE}_{ik} - \widehat{GE}_{jk}) + (e_{ik} - e_{jk}))^2,$$

and,

$$M_{ij}^2 = \left(\sum_{k=1}^{K} ((\hat{G}_i - \hat{G}_j) + (\widehat{GE}_{ik} - \widehat{GE}_{jk}) + (e_{ik} - e_{jk})) \right)^2 \Big/ K^2$$

Note that when genotypes i and j have the same interaction pattern (so $GE_{ik} = GE_{jk}$, for all k), $D_{ij}^2/2\sigma^2$ will follow a χ_{K-1}^2 distribution and hence D_{ij}^2 will have expected value $2\sigma^2(K-1)$. Thus the expected value of $I_{ij}^2 = D_{ij}^2/(K-1)$ is $2\sigma^2$ and so does not depend on the number of environments. This ensures comparability of the I_{ij}^2 interaction distance measures, from one pair of interaction similar genotypes to another, when we encounter missing values.

When genotypes i and j have the same profile, $KM_{ij}^2/2\sigma^2$ follows a non-central χ_1^2 distribution, with non-centrality parameter $\sqrt{K/2\sigma^2}$. It follows that M_{ij}^2 has expected value $2\sigma^2/K + (\hat{G}_i - \hat{G}_j)^2$. The quantity $2\sigma^2/K$ found in this expectation is generally small when compared with $(\hat{G}_i - \hat{G}_j)^2$. Thus M_{ij}^2 serves as a satisfactory measure of difference in genotype level.

In situations where there is a substantial amount of missing data, for some genotype pairs there will be little or no environment commonality. An adaptation of the strategy of Ouyang *et al.* (1995) for adjusting distances based on few comparisons and estimating unavailable distances between pairs of genotypes is used, and now described.

For any pair of genotypes we examine the number of common environments p_{ij} in which both were grown. If $p_{ij} \geqslant q$, where q is a pre-determined number, we deem the distance to have been calculated over a sufficient number of environments and use the observed distance in clustering. If $p_{ij} < q$ we must find some means of estimating the distance.

Ouyang *et al.* (1995) used the maximum value in the distance matrix to estimate unobserved distances. Their use of the maximum value in the distance matrix was justified, as the selection of genotypes on trial was based on geographic location. Distances between locations that had little commonality of genotype test sets were likely to be large. This is not necessarily the case in general. We therefore use shortest path estimates of unobserved distances in the following way. Suppose that genotypes A_i and A_j do not have any environments in common, but n other genotypes $B_1, \ldots, B_n$ share at least q environments with both A_i and A_j. We can estimate an upper bound d_{ij} for the distance between A_i and A_j as the minimum of the distances $d(A_iB_1) + d(B_1A_j), \ldots, d(A_iB_n) + d(B_nA_j)$. (The d used in this and following expressions denotes a distance function, whether it be squared Euclidean, interaction, or main effect distance.) We then use this upper bound as an estimate of the unobserved distance.

We now consider distances which are based on a number of environments less than the number we deem sufficient, so $p_{ij} < q$; we extend the Ouyang *et al.* (1995) strategy. We calculate $d(A_iA_j)$ and d_{ij} and use

$$(p_{ij}d(A_iA_j) + (q - p_{ij})d_{ij})/q$$

as the estimated distance in clustering, so combining direct and indirect information. This provides a complete set of distances, allowing clustering and then imputation to be performed. Distances I_{ij} and M_{ij} will now be used to cluster genotypes efficiently.

TWO-STAGE CLUSTERING

The presence of G × E interaction makes comparisons among genotypes a difficult task. Lin (1982), in a complete data set, removed the differences in the levels of genotypes by centring the rows of the G × E matrix, so that clusters of genotypes that performed similarly across environments could be found. Mean performances were then compared to establish which genotypes in each cluster performed best. This approach enables a researcher to reduce the number of genotypes that need to be compared, in future

testing, to the number of clusters found. On the other hand, Ivory *et al.* (1991) used the column (environment) centred yields to cluster environments, and then compared genotypes by their mean performance in each of these clusters of 'similar' environments.

In both cases the aim was to find a set of genotypes which performed similarly in a set of environments, that is, to find subsets of the original data matrix where there is no significant G × E interaction. Both approaches make analysing G × E data a simpler exercise; neither, however, is immediately capable of handling data sets which have missing entries in the G × E matrix.

We now apply the theoretical considerations of the previous section and propose the following two-stage clustering method for handling incomplete G × E matrices. This method differentiates between G × E interaction and mean performance.

First stage

(i) Calculate all interaction distances I_{ij} using (4) and the G × E matrix.
(ii) Cluster genotypes using these interaction distances. This produces clusters of shape-similar genotypes.

Second stage

(i) Calculate main effect distances M_{ij} within each first stage cluster using (3) and the G × E matrix.
(ii) Cluster genotypes within the first stage clusters using these main effect distances. This produces final clusters of level-and-shape similar genotypes.

Given dissimilarities between pairs of genotypes (in both first and second stages) a decision must be made on how clusters are to be formed, commonly referred to as the linkage method. The linkage method describes how to measure the distance between a single observation and a previously formed cluster, or between two previously formed clusters. We have chosen to use the incremental sum of squares method for forming clusters (Ward 1963). This method successively merges two current clusters so as to minimize the increase in within-clusters sum of squared distances. This method has the advantage of leading to a simple stopping criterion; we stop forming clusters when the next cluster to be formed would have an average within-cluster sum of squared distances that is greater than the average sum of squared distances in the entire set of observations. This method was used by Corsten and Denis (1990) when they simultaneously clustered genotypes and environments using their G × E interaction.

Comments on variance stabilization conclude this section. Changes in variance across environments will limit the success of this approach; we agree with the findings of Fox and Rosielle (1982) and advocate transformation within environments to equalize variance, ensuring that environments contribute similarly to distance measures. This is of particular importance in the case of missing G × E data as the genotypes need to be given an opportunity to contribute to results, irrespective of the subset of environments in which they are grown.

TWO-STAGE IMPUTATION

Often there are insufficient resources to test all genotypes in all environments, yet researchers would still like to estimate the yield from an untested combination. The knowledge gained from clustering genotypes can indicate how a genotype would perform in an environment in which it has not been tested. We shall use known similarities and observations to estimate unknown performances; the central idea is to find a substitute genotype and to use it to estimate the unknown performance. An existing strategy proposed by Drake (1981) also uses output from clustering to estimate missing observations.

Routines for the statistical package S-Plus apply the approach of Drake (1981), which we refer to as the 'nearest cluster' method. For a genotype with missing values the method uses the closest cluster of genotypes, under squared Euclidean distance (the numerator of (2)), to impute the missing values. Specifically, a genotype with a missing yield that first merges with a cluster of other genotypes will use the mean of that cluster as the estimate for the missing yield.

We now use the results from two-stage clustering to impute missing values. A step-by-step description of this process follows.

Step 1. Perform first-stage clustering, in the standard way, or, if necessary, until the cluster of the genotype with the missing value has an observation in the environment of the missing value.
Step 2. For a genotype with a missing G × E yield, identify the genotypes in the first stage cluster to which it belongs.
Step 3. Use the difference in level, between the genotype with the missing yield and every other genotype in the first stage cluster, to adjust the yield of the latter in the environment where the yield is missing.
Step 4. Calculate the mean of the estimates found in Step 3. This is the imputed value for the missing yield.
Step 5. If an imputed value is greater than the observed maximum (or less than the observed minimum) for a given environment then replace the imputed value by the observed environment maximum (minimum).

Continuation of clustering in Step 1 is necessary if we are to find imputed values for all untested

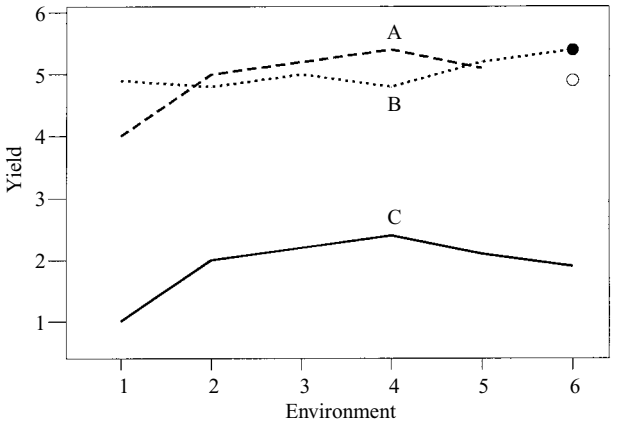

Fig. 1. Yields of three genotypes plotted against environment. Genotypes A and C are deemed similar by interaction distance, while genotypes A and B are closest using squared Euclidean distance.

combinations. The final step prevents imputation of negative yields; in fact we will not produce an imputed value that falls outside the observed range of yields in an environment, thus ensuring that there are no unrealistic imputed values. When decisions are made over selection of the best genotypes for an environment this 'trimming' guarantees that at least one tested genotype will be selected.

The difference in level mentioned in Step 3, is equal to $\bar{y}^{(j)}_{i\cdot} - \bar{y}^{(i)}_{j\cdot}$. If few or no common environments exist, then we use the intermediate genotypes $B_1, \ldots, B_n$, introduced in the Main Effect and Interaction Distances section. An overall difference in level can be approximated by the sum, over a path, of the differences in level. These overall difference estimates are then averaged. In our earlier notation this level difference is

$$\sum_{l=1}^{n} ([\bar{y}^{(B_l)}_{A_i\cdot} - \bar{y}^{(A_i)}_{B_l\cdot}] + [\bar{y}^{(A_j)}_{B_l\cdot} - \bar{y}^{(B_l)}_{A_j\cdot}])/n$$

Figure 1 illustrates the performance pattern of three genotypes across six environments; the data are artificial but it allows us to contrast the two imputation methods. Genotype A has a missing yield in the sixth environment, while in the other five environments it is most similar in shape to genotype C (a distance measure based on G × E interaction would be near zero for genotypes A and C). Measuring the difference in level between A and C and adjusting the yield of C in the sixth environment by this amount will give an estimate of the missing yield, indicated by the point marked with an open circle in Fig. 1. The imputation is therefore based on the fact that the genotypes have similar interaction profiles; it does not mix the main effect and interaction as would nearest cluster imputation.

On the other hand, a pair of genotypes which are nearest clusters, such as genotypes A and B, do not necessarily provide good substitutes for each other when values are missing. This is seen in Fig. 1, where the yield of genotype B in the sixth environment, marked with a solid circle, appears less appropriate for genotype A than the two-stage imputed value.

We conclude this section with some general comments comparing two-stage imputation with established imputation methodology. The success of two-stage imputation relies on the ability of remaining data to reconstitute truly similar interaction profiles; it is, therefore, driven by the data itself rather than by some pre-determined class of models. The ability of a class of models to adequately describe the data will be difficult to establish with incomplete data. As a consequence, there is a risk in choosing an unsuitable class of model for an incomplete data set. A second concern is the ability of certain models, when only partial data are available, to reconstitute information that would be gained from complete data. This difficulty is particularly evident when no replicates for a given genotype–environment combination are available. For these reasons, the model fitted may strongly influence the value that is imputed, for example, use of EM-AMMI (Gauch & Zobel 1990). The EM-AMMI method, or any model-based method, appears to be more useful when imputing missing replicate data, rather than when imputing missing data in unreplicated trials. In the first case a satisfactory model is more clearly determined, while in the second the two-stage imputation method is expected to be more appropriate.

AN APPLICATION

We illustrate the two-stage clustering approach, and compare two-stage imputation with nearest cluster imputation, by applying it to a data set from the G × E literature. Mungomery *et al.* (1974) first reported the experiment from which the data originat-

Table 1. *The 15 induced missing yield observations* (*standardized within environments*) *and the imputed values found using both the two-stage and the nearest cluster approaches. An asterisk marks genotypes for which the closer imputed value was found by the two-stage approach.* (*Data source: Basford & Tukey 1998*)

Genotype	Environment	Omitted yields	Two-stage imputation	Nearest cluster imputation
2*	*R*70	−0·562	−0·330	−0·218
5	*L*71	−0·648	0·868	0·799
5*	*N*71	0·497	0·996	1·301
6*	*N*71	1·230	0·551	−0·145
7	*R*71	1·434	0·612	0·948
10*	*L*71	0·852	0·923	−0·062
14	*N*70	0·930	0·141	0·469
19*	*B*70	−1·308	−0·847	−0·121
19	*L*71	0·375	−1·282	−1·263
24	*B*70	0·809	−0·563	−0·557
26	*B*71	0·738	−0·291	0·039
30*	*N*70	−0·480	−0·174	−0·082
37*	*N*71	−0·599	−0·093	−0·007
52*	*B*70	−1·483	−0·939	−0·841
53*	*R*70	1·674	1·625	1·469

ed. It has been analysed in many different and sometimes innovative ways, including Basford (1982) and Basford & McLachlan (1985). Basford & Tukey (1998) published the data set in full.

Six response variables were evaluated for 58 soybean genotypes from four locations over two years. Using the location-year combination as the environment gives a G × E matrix that is 58 × 8 in size with entries being mean yields of two replicates from a randomized block design. As we do not want the variation within environments to affect the analysis in this illustrative example, the yields have first been standardized within each environment, using the transformation

$$z_{ik} = \frac{y_{ik} - \bar{y}_{\cdot k}}{s_k}$$

where $\bar{y}_{\cdot k}$ and s_k are the kth environment mean and standard deviation, respectively. This transformation does not alter the qualitative structure of the G × E interaction that exists in the data, but ensures that each environment contributes on an equal footing to the distance measures calculated.

We present an example that has 15 points randomly removed from the G × E matrix and then impute these using two-stage clustering-based imputation and the nearest cluster method. The deleted observations and their imputed values are listed in Table 1. We now illustrate how imputation is performed by each method.

Figure 2 presents the clustering of genotypes using the mean squared Euclidean distance from Ouyang *et al.* (1995) in (2) and the incremental sum of squares method of forming clusters. Use of this averaged distance measure improves on the original implementation of Drake (1981), being more appropriate for the reasons discussed in the Main Effect and Interaction Distances section. Applying nearest cluster imputation would, for instance, use the yields of genotypes 51 and 52 in place of each other's missing yields as these are deemed closest to each other. Thus the missing yield of genotype 52 in environment B70 would be imputed using the yield of genotype 51 in B70, namely −0·841.

Figure 3 shows the first stage dendrogram for the 58 genotypes clustered using interaction distance I_{ij}, appropriate when data are incomplete, and the incremental sum of squares method. The similarity of the clustering in Figs 2 and 3 reflects the low importance of the differences in level between many of the genotypes under examination. An exception to this is genotype 54, which is clustered differently in the two figures; its level similarity with genotypes 8 and 29 forces the clustering in Fig. 2, while its distinct G × E interaction profile forces it to remain outside the clustering until late in Fig. 3.

First stage clustering was stopped when 10 clusters remain, at level 196, determined using the stopping criterion described in the section on clustering presented earlier. For illustrative purposes we focus on one cluster of interaction-similar genotypes, seen at the left of Fig. 3, containing genotypes 14, 1, 28, 23, 31, 40, 15, 35, 37, 34, 38 and 39.

Figure 4 then shows the second stage dendrogram for this first stage cluster, using main effect distance M_{ij} appropriate for incomplete data, and the incremental sum of squares linkage method. As the previously selected cluster is a little large, we now use a smaller first stage cluster to illustrate the imputation of a missing yield. We use the first stage cluster that

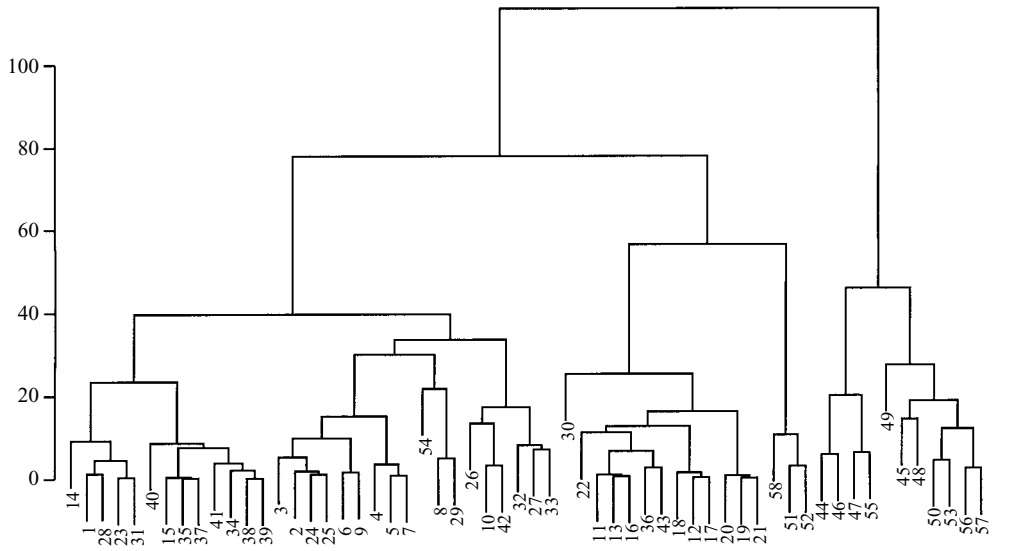

Fig. 2. Fifty-eight genotypes clustered using the mean squared Euclidean distance of Ouyang *et al*. (1995) and incremental sum of squares.

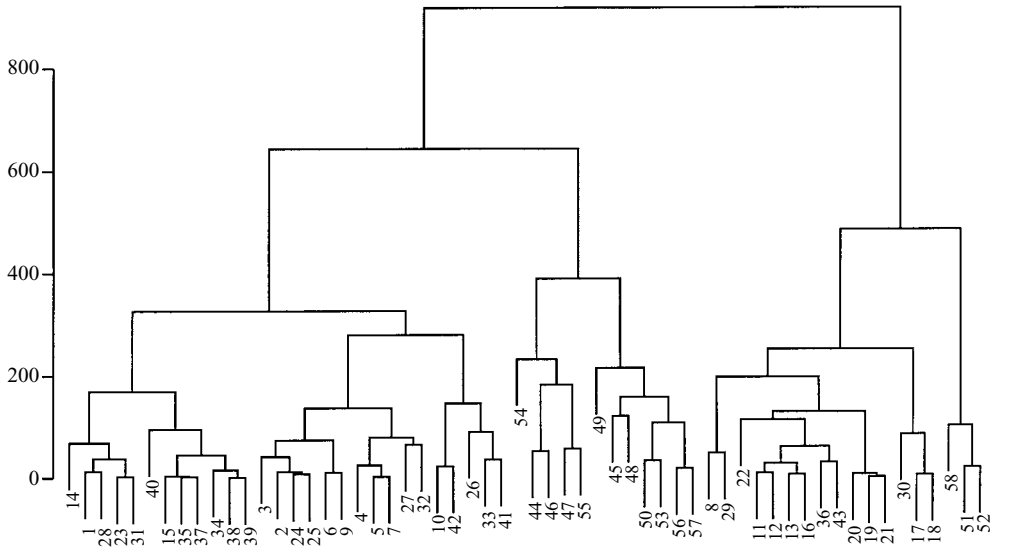

Fig. 3. First stage clustering: 58 genotypes clustered by similarity of interaction profile, using interaction distance I_{ij} and the incremental sum of squares method.

contains genotypes 51, 52 and 58 to illustrate the imputation of the missing value of genotype 52 in B70. The yields of these three genotypes are plotted against environments in Fig. 5.

Two-stage imputation of a missing yield does not use the results of second stage clustering but does use the second stage distance. Recall that first stage clustering produces clusters of genotypes that have similar interaction profiles; differences in level between these genotypes are then used to impute a missing value. Thus, for example, the missing yield of genotype 52 in B70 will be imputed using genotypes 51 and 58, as these are the only genotypes deemed similar to genotype 52 when first stage clustering is truncated. In B70, genotype 51 yields $-0{\cdot}841$, while genotype 58 yields $-1{\cdot}278$. Genotype 51 on average yields 0·250 less than genotype 52, so provides an estimate of the standardized yield for genotype 52 in

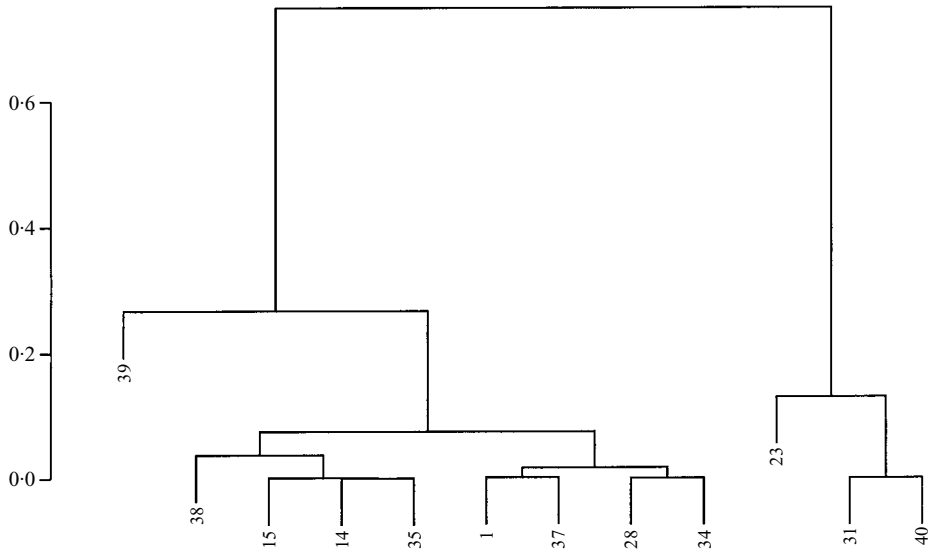

Fig. 4. Second stage clustering: 12 similar genotypes, a first stage cluster seen towards the left of Fig. 3, clustered using main effect distance M_{ij} and the incremental sum of squares method.

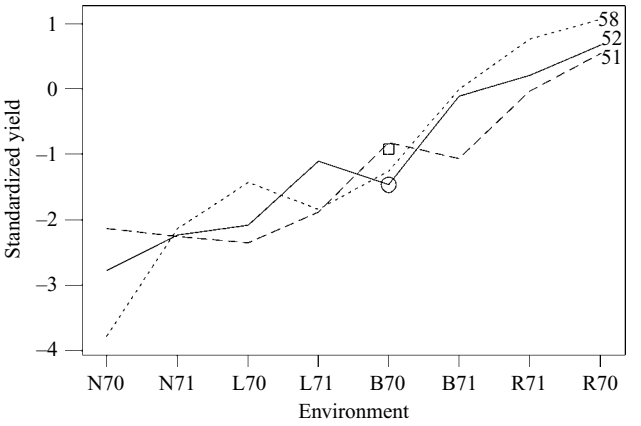

Fig. 5. Yields of the three similar genotypes (51, 52 and 58) plotted against an ordered environmental index, calculated as the mean of these genotypes using imputed values where necessary. The point marked with an open square is the imputed value of genotype 52, and the open circle marks the omitted value.

B70 of −0·591. Similarly, genotype 58 on average yields 0·009 more than genotype 52, giving an estimate of −1·287. These two values are then averaged to give the imputed value for genotype 52 in B70 of −0·939, found in Table 1.

Figure 5 reproduces the yield profiles of the three genotypes mentioned in the imputation of genotype 52 and indicates the omitted yield of genotype 52 in B70. Genotypes 51 and 58 are similar in their interaction profiles to genotype 52 as determined by first stage clustering. The imputed value for genotype 52, found using the two-stage approach, is marked with an open square in Fig. 5, while an open circle marks the omitted yield.

It can be seen from Table 1 that in 9 of the 15 cases the two-stage method provides a closer imputed value than the nearest cluster approach. This is typical for this amount of missing data, as will be shown in the next section.

For this example, an overall comparison of the results from the two imputation methods, using mean squared error (MSE) of the estimate, shows that the two-stage imputation method performs better, with an MSE of 0·727 compared with an MSE of 0·896 for the nearest cluster method. The next section compares the methods by varying the amount of missing data in the G × E matrix, using a number of G × E data sets available in the literature.

COMPREHENSIVE TESTING

Comprehensive testing of two-stage imputation is required to compare its effectiveness with the nearest cluster method. Both these methods are also compared with imputation using values randomly selected from the observed yields of other genotypes within the same environment. Varying amounts of missing data have been simulated by randomly deleting values from complete G × E matrices in order to compare the two imputation methods. The following procedure has been used on a variety of data sets from the G × E literature:

1. Randomly remove the desired number of elements from the complete G × E matrix.
2. Check this new matrix for the representation of each genotype and environment. If each genotype (environment) is not represented in some minimum number of environments (genotypes), then start again.
3. If this matrix is partitioned, that is, there is no direct or indirect commonality of environments between every pair of genotypes, then start again.
4. Impute missing values using all methods.
5. Record the mean squared error for all methods, and for each pair of methods, record the proportion of cases imputed more accurately by each method.

We now add greater detail to the above summary.

We have tested the effectiveness of two-stage imputation using five data sets; the Mungomery *et al.* (1974) data described in the previous section, data sets from Gauch (1992), Ramey & Rosielle (1983), and two from Flores *et al.* (1998) have been used as they are moderately large, and more importantly, complete. Table 2 shows the size and shape of the G × E matrices and the G × E interaction sum of squares. Apart from the Mungomery *et al.* (1974) set, we have not transformed the data, but have used the means over G × E combination replicates. We believe that appropriate transformation should be considered as a general rule to equalize the within-environment variance.

Two approaches have been used to ensure that sufficient information remains in the G × E matrix after data removal. First, each incomplete matrix has been checked to ensure that all genotypes were grown in a minimum number of environments. This minimum level was chosen at four environments in our testing, as this was at least half of the total number of environments in the smaller data sets used. Second, we ensured that the incomplete G × E matrix formed by data deletion did not lead to two unlinked data sets, so that it remains possible to impute all missing G × E yields.

Imputation via either the two-stage or nearest cluster method, for some G × E combinations, will be impossible if we allow the G × E matrix to become unlinked. We have allowed, however, a pair of genotypes to have no common environments. We use the strategy outlined in the Main Effect and Interaction Distances section presented earlier in this instance. We have set the value of q to be four in our work, as this integer must be less than or equal to the minimum representation of environments in a given G × E matrix. If we allowed q to exceed the minimum representation, an under-represented genotype would have undefined distances to every other genotype.

The imputations are then compared with the deleted values. We have compared any two methods using both the MSE of the imputed values and the proportion of values that were imputed more accurately by one method than the other. For comparative purposes we have also used a value randomly selected from those within the environment, as a third imputation method. We have performed 1000 runs for each data set and missing data amount so that a sufficiently large number of randomly imputed values can be compared with imputed values found by the two-stage and nearest cluster methods. These results are shown in Table 3. We note that it is possible for two methods to give the same imputed value; this is counted in favour of the second method in any comparison and thus the figures provided understate the performance of the first method.

The two-stage method consistently outperforms the nearest cluster method; this margin increases as the amount of missing data increases, in the sets that have a large number of environments. The reason for this appears to be the ability of the reduced data set to

Table 2. *Summary details of the data sets used in testing. Note that the Mungomery data are standardized and that this has an effect on the magnitude of the genotype main effect and interaction sums of squares*

Dataset	Number of genotypes	Number of environments	Genotype SS	G × E interaction SS	G × E interaction SS as % of total SS
Mungomery	58	8	238·44	217·56	47·71
Ramey	15	9	9506462	16334640	63·21
Gauch	7	10	7117668	39728718	84·81
Flores 1	15	12	17120782	18191558	51·52
Flores 2	11	16	5634380	24798300	81·49

Table 3. *Comparison of all pairs of methods using five data sets and varying levels of missing data. The two measures are the percentage of 1000 runs for which the mean squared error for the first method is lower, and (in parentheses) the average proportion of the first method's imputed values that are closer to the missing values*

Dataset	Number of points removed	Two-stage vs. nearest cluster	Two-stage vs. random	Nearest cluster vs. random
Flores 1	3	76·7 (0·627)	79·7 (0·664)	64·0 (0·567)
	5	80·0 (0·610)	84·1 (0·646)	72·2 (0·566)
	10	87·6 (0·606)	92·3 (0·648)	79·6 (0·562)
	15	89·7 (0·596)	95·4 (0·637)	85·0 (0·552)
	20	92·9 (0·590)	97·2 (0·630)	88·9 (0·551)
	25	92·5 (0·583)	98·3 (0·631)	90·4 (0·549)
	30	92·8 (0·580)	98·8 (0·634)	92·8 (0·554)
	35	89·3 (0·569)	99·1 (0·624)	93·9 (0·548)
Flores 2	3	66·2 (0·562)	66·3 (0·606)	60·0 (0·549)
	5	67·6 (0·545)	71·2 (0·598)	62·4 (0·542)
	10	70·1 (0·541)	72·4 (0·583)	64·6 (0·527)
	15	69·3 (0·523)	73·6 (0·576)	65·3 (0·520)
	20	75·4 (0·529)	75·8 (0·576)	64·3 (0·513)
	25	74·4 (0·523)	79·3 (0·576)	68·9 (0·512)
	30	77·2 (0·522)	80·0 (0·573)	67·2 (0·510)
	35	77·8 (0·522)	79·2 (0·564)	65·6 (0·498)
	40	79·3 (0·517)	79·4 (0·561)	64·3 (0·493)
	45	79·5 (0·516)	79·8 (0·557)	64·4 (0·492)
	50	81·3 (0·513)	79·4 (0·552)	65·6 (0·485)
	55	80·0 (0·508)	82·3 (0·551)	63·4 (0·486)
	60	81·0 (0·500)	80·9 (0·544)	64·8 (0·480)
Gauch	3	82·2 (0·667)	91·6 (0·710)	71·2 (0·484)
	5	77·6 (0·601)	95·0 (0·670)	82·7 (0·477)
	10	82·3 (0·636)	92·3 (0·679)	81·8 (0·501)
Mungomery	3	64·3 (0·586)	85·9 (0·696)	77·5 (0·644)
	5	70·1 (0·597)	90·0 (0·702)	82·0 (0·648)
	10	77·4 (0·598)	97·0 (0·701)	90·3 (0·635)
	15	84·9 (0·598)	98·3 (0·699)	92·4 (0·631)
	20	86·0 (0·588)	99·6 (0·707)	97·3 (0·640)
Ramey	3	59·6 (0·522)	75·6 (0·633)	65·8 (0·574)
	5	63·7 (0·514)	77·8 (0·620)	69·5 (0·573)
	10	71·6 (0·509)	84·4 (0·622)	73·8 (0·567)
	15	72·8 (0·507)	86·5 (0·615)	74·9 (0·567)

retain the qualitative G × E interaction structure of the complete matrix when there is more information. In such cases the clustering of similar genotypes is less likely to change as data are deleted. The improvement of the two-stage method over the nearest cluster method is always greater than in the case when only three points were removed from the Ramey & Rosielle (1983) data; the mean squared error in this worst case was lower in only 59·6 % of all runs. When the second criterion is considered, the two-stage imputed values are consistently more likely to be closer to the omitted value than the nearest cluster-imputed values. The case where 60 points were removed from the second data set from Flores *et al.* (1998) had the lowest average proportion (0·500) of omitted values imputed better by two-stage imputation. The mean squared error comparison for this worst case, however, shows that while on average only half of the omitted values were imputed more accurately using two-stage imputation, the MSE was lower for two-stage imputation in 81·0 % of runs. The advantage gained by the two-stage imputed values, when they were closer to the omitted yields than nearest cluster-imputed values, exceeds that gained by the nearest cluster-imputed values when they are closer to the omitted yields.

The nearest cluster method does not always outperform random imputation in terms of the proportion of imputed values that are closer to the omitted yield. On the other hand, when the criterion used to gauge effectiveness is the MSE, the nearest cluster method consistently outperforms the use of random values. Based on MSE, the two-stage method always improves on random insertion more strongly than does the nearest cluster method.

No discernible trends with the G × E interaction are identified in the results. This is contrary to our initial

belief that a method which separates the interaction component would increase in effectiveness as the relative size of the G × E interaction increases. Identification of other reasons for the success of the two-stage method will require further simulation testing with more data sets and possibly greater amounts of missing data.

CONCLUSION

Squared Euclidean distance between a pair of genotypes can be partitioned naturally into two orthogonal and independent components, one representing interaction of genotype and environment and the other a main effect, or difference in level. These two components lead to the two stages of the clustering method presented, appropriate for handling missing data in genotype-by-environment studies. First we cluster genotypes by interaction similarity, then we cluster similar genotypes by their main effect.

We provide evidence that this partitioning of squared Euclidean distance also leads to an improved method for imputing data in genotype-by-environment analyses. The imputation method based on two-stage clustering outperforms the nearest cluster-based method, using two criteria. Furthermore, when judged by mean squared error, it becomes more effective as the amount of missing data in the G × E matrix increases.

The authors wish to extend their greatest thanks to Dr Lesley Currah for providing the motivation for this work. The authors are grateful to Zoë Wood for assistance with the preparation of this paper, and to referees, whose comments led to considerable improvement to the paper.

REFERENCES

Basford, K. E. (1982). The use of multidimensional scaling in analysing multi-attribute genotype response across environments. *Australian Journal of Agricultural Research* **33**, 473–480.

Basford, K. E. & Mclachlan, G. J. (1985). The mixture method of clustering applied to three-way data. *Journal of Classification* **2**, 109–125.

Basford, K. E. & Tukey, J. W. (1998). *Graphical Approaches to Multiresponse Data: Illustrated with a Plant Breeding Trial*. London: Chapman and Hall.

Cooper, M. & Delacy, I. H. (1994). Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. *Theoretical and Applied Genetics* **88**, 561–572.

Corsten, L. C. A. & Denis, J. B. (1990). Structuring interaction in two-way tables by clustering. *Biometrics* **46**, 207–215.

Crossa, J. (1990). Statistical analyses of multilocation trials. *Advances in Agronomy* **44**, 55–85.

Drake, D. W. (1981). *The GEBEI analysis package*. Queensland Department of Primary Industries, Brisbane. Misc. Pub. 81022.

Flores, F., Moreno, M. T. & Cubero, J. I. (1998). A comparison of univariate and multivariate methods to analyse G × E interaction. *Field Crops Research* **56**, 271–286.

Fox, P. N. & Rosielle, A. A. (1982). Reducing the influence of environmental main-effects on pattern analysis of plant breeding environments. *Euphytica* **31**, 645–656.

Freeman, G. H. (1973). Statistical methods for the analysis of genotype-environment interactions. *Heredity* **31**, 339–354.

Freeman, G. H. (1985). The analysis and interpretation of interactions. *Journal of Applied Statistics* **12**, 3–9.

Gauch, H. G. (1992). *Statistical Analysis of Regional Yield Trials: AMMI Analysis of Factorial Designs*. Amsterdam, The Netherlands: Elsevier Science Publishers B. V.

Gauch, H. G. & Zobel, R. W. (1990). Imputing missing yield trial data. *Theoretical and Applied Genetics* **79**, 753–761.

Godfrey, A. J. R., Wood, G. R. & Nichols, M. A. (1999). Gotta know your onions, or, a new approach for clustering cultivars in genotype-by-environment analysis with sparsity in the data. In *Proceedings of the New Zealand Statistical Association 50th Anniversary Conference* pp. 36–37. Wellington, New Zealand.

Ivory, D. A., Kaewmeechai, S., Delacy, I. H. & Basford, K. E. (1991). Analysis of the environmental component of genotype × environment interaction in crop adaptation evaluation. *Field Crops Research* **28**, 71–84.

Lin, C. S. (1982). Grouping genotypes by a clustering method directly related to genotype-environment interaction mean square. *Theoretical and Applied Genetics* **62**, 277–280.

Lin, C. S., Binns, M. R. & Lefkovitch, L. P. (1986). Stability analysis: Where do we stand? *Crop Science* **26**, 894–900.

Mungomery, V. E., Shorter, R. & Byth, D. E. (1974). Genotype × environment interactions and environmental adaptation. I: Pattern analysis – application to soya bean populations. *Australian Journal of Agricultural Research* **25**, 59–72.

Ouyang, Z., Mowers, R. P., Jensen, A., Wang, S. & Zheng, S. (1995). Cluster analysis for genotype × environment interaction with unbalanced data. *Crop Science* **35**, 1300–1305.

Ramey, T. B. & Rosielle, A. A. (1983). HASS cluster analysis: A new method of grouping genotypes or environments in plant breeding. *Theoretical and Applied Genetics* **66**, 131–133.

Ward, J. H. (1963). Hierarchical grouping to optimise an objective function. *Journal of the American Statistical Association* **58**, 236–244.

Two-stage clustering in genotype-by-environment analyses with missing data

Godfrey, A. J. R.

2002